SUT Journal of Mathematics
Vol. 42, No. 1 (2006), 123–131

Fixed width confidence interval for equal means with intraclass correlation model

Hiroto Hyakutake, Masaaki Anan and Tomoya Mizuyoshi

(Received January 12, 2006; Revised June 27, 2006)

Dedicated to Professor Minoru Siotani on his 80th birthday

Abstract. In repeated measures, p dimensional measurements are observed, then testing a hypothesis for equality of means of p components is one of the simplest inference. If the hypothesis were accepted, it is interesting to estimate of the mean. In this paper, we consider the problem of constructing a fixed width confidence interval of equal normal means, when the covariance matrix is intraclass correlation model.

AMS 2000 *Mathematics Subject Classification.* 62F25.

Key words and phrases. Confidence interval, intraclass correlation model, missing data, two-stage procedure.

§1. Introduction

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ be independent and identically distributed random vectors having p-variate normal distribution with mean $\mu\mathbf{1}_p$ and covariance matrix Σ, that is $N_p(\mu\mathbf{1}_p, \Sigma)$, where $\mathbf{1}_p$ is a p dimensional vector of ones. We assume that the covariance matrix have the structure $\Sigma = \sigma^2\{(1-\rho)I_p + \rho\mathbf{1}_p\mathbf{1}_p'\}$, which is called an intraclass correlation model, where $-1/(p-1) < \rho < 1$. Let $\bar{x}_n = \sum_{i=1}^n \mathbf{1}_p'\boldsymbol{x}_i/np$. The distribution of $\bar{x}_n$ is $N(\mu, \tau/np)$, where $\tau = \sigma^2\{1+(p-1)\rho\}$, which is a characteristic root of Σ. The problem is to determine the sample size satisfying

$$P\{|\bar{x}_n - \mu| \le d\} \ge 1-\alpha, \tag{1.1}$$

where $d > 0$ and α $(0 < \alpha < 1)$ are given. The left hand side of (1.1) is $2\Phi(d/\sqrt{\tau/np}) - 1$, where Φ is the cumulative distribution function of $N(0,1)$. It is easily seen that if σ^2 and ρ were known and the sample size n were determined such that

$$n \ge n^* = z_{\alpha/2}^2\tau/pd^2, \tag{1.2}$$

where $z_{\alpha/2}$ is the upper $100\alpha/2\%$ point of $N(0,1)$, then (1.1) is satisfied.

If σ^2 and ρ are unknown, no fixed sample size procedure exists to achieve (1.1). We propose a two-stage procedure to achieve (1.1). The two-stage procedure is originally proposed by Stein [6]. Hyakutake, Takada and Aoshima [2] proposes a two-stage procedure, when the covariance matrix is the intraclass correlation model and the components of mean vector are not equal, say $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)'$. Testing a hypothesis $\mu_1 = \cdots = \mu_p$ is reviewed in Siotani, Hayakawa and Fujikoshi [3].

In this paper, we give a two-stage procedure satisfying (1.1) in Section 2. We consider the problem (1.1) with missing data in Section 3. The problem of missing data arises in many situation, for example repeated measurement analysis, familial data analysis, and so on, see Srivastava [5]. It is not easy to give an exact procedure, so approximated procedures are proposed. In Section 4, the accuracy of approximation is examined by simulation and a numerical example by using cholesterol data in Wei and Lachin [7] is given.

§2. Two-stage procedure

It is well known that there exists no fixed width confidence interval for the mean with the fixed sample size when the variance is unknown. In this section, we propose a two-stage procedure satisfying (1.1). Healy [1] proposed a multivariate two-stage procedure, which is an extension of Stein's [6] univariate two-stage sampling scheme. In the sampling rule of the two-stage procedure, one takes samples of size $m(> p)$ and compute the sample covariance matrix

$$S_m = \frac{1}{m-1}\sum_{i=1}^{m}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$$

and $\hat{\tau} = \mathbf{1}_p' S_m \mathbf{1}_p/p$, where $\bar{\boldsymbol{x}} = \sum_{i=1}^{m} \boldsymbol{x}_i/m$. Then the total sample size N is defined by

$$N = \max\{m,\ [\hat{\tau}t^2/pd^2] + 1\}, \tag{2.1}$$

where t is the upper $100\alpha/2\%$ point of the t-distribution with $m-1$ degrees of freedom (d.f.) and $[a]$ denotes the greatest integer not greater than a. Note that $(m-1)\hat{\tau}/\tau$ has a chi-square distribution with $m-1$ d.f. (χ^2_{m-1}), see e.g. Srivastava [4]. Next take $N-m$ additional observations and compute the sample mean $\bar{x}_N$. Since $N \geq \hat{\tau}t^2/pd^2$ by (2.1) and $(\bar{x}_N - \mu)/\sqrt{\hat{\tau}/Np}$ has the

t-distribution with $m-1$ d.f.,

$$\begin{aligned} & P\{|\bar{x}_N - \mu| \le d\} \\ \ge\ & P\{|\sqrt{N}(\bar{x}_N - \mu)|/\sqrt{\hat{\tau}t^2/pd^2} \le d\} \\ =\ & P\{|\bar{x}_N - \mu|/\sqrt{\hat{\tau}/Np} \le t\} \\ =\ & 1-\alpha. \end{aligned}$$

Hence (1.1) is satisfied.

§3. Missing observations

In this section, the problem (1.1) is considered, when the observations are monotone type of missing. Let the ith observation be $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip_i})'$ $(1 \le p_i \le p)$, say the last $p - p_i$ components are missing. Then the distribution of $\boldsymbol{x}_i$ is $N_{p_i}(\mu \mathbf{1}_{p_i}, \sigma^2\{(1-\rho)I_{p_i} + \rho \mathbf{1}_{p_i}\mathbf{1}'_{p_i}\})$. Let $\bar{x}_n = \sum_{i=1}^n \mathbf{1}'_{p_i}\boldsymbol{x}_i/k_n$, where

$$k_n = \sum_{i=1}^n p_i \tag{3.1}$$

is the sum of the number of observed components, then the distribution of $\bar{x}_n$ is

$$N\Big(\mu, \frac{\sum_{i=1}^n p_i\xi_i}{k_n^2}\Big), \tag{3.2}$$

where $\xi_i = \sigma^2\{1 + (p_i - 1)\rho\}$. Note that $\bar{x}_n$ is not the maximum likelihood estimator of μ. In (1.1), the equality holds when $n = n^*$ and data has no missing. If we define the sample size as (1.2), then (1.1) will not be satisfied with missing data. The variance in (3.2) depends on p_i, hence the required sample size depends on the observed dimension. So, we wish to give k_n defined in (3.1) in order to satisfy (1.1).

If $\rho \ge 0$, then $\sum_{i=1}^n p_i\xi_i/k_n^2 \le \tau/k_n$ by $\xi_i \le \tau$. Hence (1.1) is satisfied when $k_n \ge z_{\alpha/2}^2\tau/d^2$ for $\rho \ge 0$. This is same as (1.2) when $p_i = p$ $(i = 1, \cdots, n)$. If $\rho < 0$, then $\sum_{i=1}^n p_i\xi_i/k_n^2 \le \sigma^2/k_n$. When $k_n \ge z_{\alpha/2}^2\sigma^2/d^2$ for $\rho < 0$, (1.1) is satisfied. This is equivalent to the case when $p_i = 1$ $(i = 1, \cdots, n)$, say all samples are observed only one dimension. Hence if σ^2 and ρ were known and k_n were chosen such that

$$k_n \ge k_{n^*} = z_{\alpha/2}^2 \max\{\tau, \sigma^2\}/d^2, \tag{3.3}$$

then (1.1) is satisfied. The lower bound k_{n^*} may not be optimal, the optimal sample size $k_n = \sum_{i=1}^n p_i$ is a solution of

$$\frac{\sum_{i=1}^n p_i\sigma^2\{1 + (p_i - 1)\rho\}}{(\sum_{i=1}^n p_i)^2} = \frac{d^2}{z_{\alpha/2}^2}.$$

But it is difficult to give the explicit solution to this equation.

When σ^2 and ρ are unknown, we consider two-stage procedures. But it is not easy to give an exact procedure. The proposed procedures will be approximately satisfied (1.1). In the procedure, we take samples of size m and compute

$$v_b = \sum_{i=1}^{m} p_i(\bar{x}^{(i)} - \bar{x}_m)^2, \tag{3.4}$$

where $\bar{x}^{(i)} = \mathbf{1}'_{p_i}\boldsymbol{x}_i/p_i$. In analysis of variance, since $\sum_{ij}(x_{ij} - \bar{x}_m)^2 = v_b + v_w$, v_b and v_w are considered as the between group sum of squares and the within group sum of squares, respectively, where $v_w = \sum_{i,j}(x_{ij} - \bar{x}^{(i)})^2$. Noting that $E(v_b) = \sum_{i=1}^{m}(1 - p_i/k_m)\xi_i$ and $E(v_w) = (k_m - m)\sigma^2(1-\rho)$,

$$\hat{\sigma}^2 = \frac{v_b + \sum_{i=1}^{m}\{(1 - p_i/k_m)(p_i - 1)/(k_m - m)\}v_w}{\sum_{i=1}^{m}(1 - p_i/k_m)p_i},$$

is an unbiased estimator of σ^2. Let an estimator of ρ be

$$\hat{\rho} = 1 - v_w/\{(k_m - m)\hat{\sigma}^2\},$$

then an estimator of τ is $\hat{\tau} = \hat{\sigma}^2\{1 + (p-1)\hat{\rho}\}$.

First, we consider a two-stage procedure by asymptotic approximation. Since the estimators $\hat{\tau}$ and $\hat{\sigma}^2$ are consistent, (1.1) would be satisfied asymptotically, when $k_n \geq z^2_{\alpha/2}\max\{\hat{\tau}, \hat{\sigma}^2\}/d^2$ by (3.3). So, we define

$$\tilde{k}_{N^*} = \max\{m,\ [z^2_{\alpha/2}\max(\hat{\tau}, \hat{\sigma}^2)/d^2] + 1\} \tag{3.5}$$

and take additional samples until satisfying $k_N \geq \tilde{k}_{N^*}$. By computing $\bar{x}_N$, an approximated confidence interval is given by

$$\bar{x}_N \pm z_{\alpha/2}\sqrt{\max(\hat{\tau}, \hat{\sigma}^2)/k_N}. \tag{3.6}$$

But the coverage probability of this confidence interval is sometimes less than $1-\alpha$ in simulation given in the next section. Hence the coverage probability of the confidence interval $\bar{x}_N \pm z_{\alpha/2}\sqrt{k_N}\sqrt{\sum_{i=1}^{m} p_i\hat{\xi}_i/k_m}$ would be less than predetermined confidence coefficient, because $\sum_{i=1}^{m} p_i\hat{\xi}_i/k_m \leq \max(\hat{\tau}, \hat{\sigma}^2)$, where $\hat{\xi}_i = \hat{\sigma}^2\{1 + (p_i - 1)\hat{\rho}\}$. In (3.5), we use $z_{\alpha/2}$, because it is difficult to derive the exact distribution of $\max(\hat{\tau}, \hat{\sigma}^2)$.

Next we consider another procedure. The distribution of $\bar{x}^{(i)} - \mu$ is $N(0, \xi_i/p_i)$, so the statistic v_b in (3.4) can be expressed by using chi-square random variables as follows:

$$\begin{aligned} v_b &= \sum_{i=1}^{m} p_i(\bar{x}^{(i)} - \mu)^2 - \sum_{i=1}^{m} p_i(\bar{x}_m - \mu)^2 \\ &= \sum_{i=1}^{m} \xi_i(v_i - p_i v_0/\textstyle\sum_{i=1}^{m} p_i), \end{aligned}$$

where v_i has χ^2_1 $(i = 0, 1, \cdots, m)$. The statistics $v_1, \cdots, v_m$ are independent, but v_0 is not independent to $v_1, \cdots, v_m$. If $p_i = p$ for all i, then v_b/τ is distributed as χ^2_{m-1}. So, $u = v_b/(m-1)$ would be used for determining the total number of components. But $E(u) \leq \sum_{i=1}^m p_i\xi_i/k_m$ by

$$k_m \sum_{i=1}^m \xi_i \leq m \sum_{i=1}^m p_i\xi_i, \tag{3.7}$$

which can be obtained by $(\sum_{i=1}^m p_i)^2 \leq m \sum_{i=1}^m p_i^2$, where $\sum_{i=1}^m p_i\xi_i/k_m$ is the variance of $\sqrt{k_m}(\bar{x}_m - \mu)$. If u is used for determining k_N, the coverage probability may be also less than $1-\alpha$. Let the probability that the number (dimension) of observed components of each individual is j be θ_j, that is $P(p_i = j) = \theta_j$ $(j = 1, \cdots p)$, where $0 \leq \theta_j \leq 1$ and $\theta_1 + \cdots + \theta_p = 1$. Then $E(p_i) = \sum_{j=1}^p j\theta_j$ and $E(p_i^2) = \sum_{j=1}^p j^2\theta_j$. Let $E(p_i) = q_1$ and $E(p_i^2) = q_2$. Then we have $k_m/m = \sum_{i=1}^m p_i/m \to q_1$ and $\sum_{i=1}^m p_i^2/m \to q_2$ as $m \to \infty$. Hence if k_N is defined by a two-stage procedure, then

$$\sum_{i=1}^m p_i\xi_i/k_m \approx \sum_{i=1}^N p_i\xi_i/k_N \tag{3.8}$$

for large m. By (3.7) and $k_m^2 \leq m \sum_{i=1}^m p_i^2$, it is easy to see that $k_m \sum_{i=1}^m p_i\xi_i \leq \sum_{i=1}^m p_i^2 \sum_{i=1}^m \xi_i$. Hence we have

$$\frac{\sum_{i=1}^m \xi_i v_i}{\sum_{i=1}^m p_i\xi_i/k_m} \geq \frac{1}{\ell_m} \sum_{i=1}^m v_i \tag{3.9}$$

by $m \sum_{i=1}^m \xi_i v_i \geq \sum_{i=1}^m \xi_i \sum_{i=1}^m v_i$, where $\ell_m = m \sum_{i=1}^m p_i^2/k_m^2$. Note that $\ell_m \geq 1$, the equality holds when $p_1 = \cdots = p_m$. ℓ_m is maximized when $p_1 = \cdots p_{i'} = p$, $p_{i'+1} = \cdots = p_m = 1$, then $\ell_m = m(i'p^2 + m - i')/(i'p + m - i')^2$, where $i' = [m/(p+1)]$ or $[m/(p+1)] + 1$. Hence we have $\ell_m \leq (p+1)^2/4p$.

By (3.9),

$$\ell_m v_b \geq (\sum_{i=1}^m p_i\xi_i/k_m)(\sum_{i=1}^m v_i - \ell_m v_0). \tag{3.10}$$

Now we have the following theorem.

Theorem. Let

$$k_{N^*} = \max\{m, [\ell_m t'^2 u/d^2] + 1\} \tag{3.11}$$

and take additional samples until satisfying $k_N \geq k_{N^*}$, where t' is the upper $100\alpha/2\%$ point of the t-distribution with $m - \ell_m$ d.f. Then the confidence interval is given by

$$\bar{x}_N \pm t' \sqrt{\ell_m u/k_N}, \tag{3.12}$$

whose the coverage probability is $1-\alpha$ approximately and $t'\sqrt{\ell_m u/k_N} \leq d$.

Proof. By (3.10),

$$\begin{aligned} & P\{|\bar{x}_N - \mu| \le t'\sqrt{\ell_m u/k_N}\} \\ \ge\ & P\left\{|\bar{x}_N - \mu| \le t'\sqrt{\frac{\sum_{i=1}^m v_i - \ell_m v_0}{k_N(m-1)}\frac{\sum_{i=1}^m p_i\xi_i}{k_m}}\right\} \end{aligned}$$

This is approximated by

$$P\left\{\frac{|\bar{x}_N - \mu|}{\sqrt{\sum_{i=1}^N p_i\xi_i/k_N^2}} \le t'\sqrt{\frac{\sum_{i=1}^m v_i - \ell_m v_0}{m-1}}\right\},$$

by (3.8), where $z = (\bar{x}_N - \mu)/\sqrt{\sum_{i=1}^N p_i\xi_i/k_N^2}$ is a standard normal variable. The statistic $\sum_{i=1}^m v_i - \ell_m v_0$ would be approximated by $\chi^2_{m-\ell_m}$ variable. Hence the distribution of $z/(\sqrt{\sum_{i=1}^m v_i - \ell_m v_0})/(m-\ell_m)$ can be approximated by t-distribution with $m - \ell_m$ d.f., then the coverage probability is approximated by $1-\alpha$. It is easy to see that the length of the interval (3.12) is not greater than $2d$ by (3.11).

If there is no missing, say $p_i = p$ $(i = 1, 2, \cdots)$, the two-stage procedure (3.11) is equivalent to (2.1). Since $1 \le \ell_m \le 2$ for $p \le 5$, t' would be approximated by t when p is small.

§4. Simulation and example

In this section, we examine the accuracy of approximation of the proposed procedure and give a numerical example.

4.1. Simulation

The two-stage procedures given in the previous section are approximated procedure. We examine the accuracy by simulation. In the simulation, we choose $p = 4$, the parameters as $\mu = 5.0$, $\sigma^2 = 1$, and $\rho = 0.2,\ 0.5,\ 0.8$, predetermined constants as $d = 0.5$ and $\alpha = 0.05$, and the missing rate $\epsilon = 0.2,\ 0.4$. For these values, we construct 10,000 confidence intervals (3.6) and calculate the proportion that intervals include the true mean $\mu = 5.0$. The results are in Table 1, in which some of values are smaller than $1 - \alpha = 0.95$. Particularly, the values for $\rho = 0.2$ are significantly smaller than 0.95.

Table 1. Coverage probability of (3.6)

ϵ	ρ		
	0.2	0.5	0.8
0.2	0.9448	0.9466	0.9471
0.4	0.9429	0.9584	0.9599

Next, we construct 10,000 confidence intervals (3.12), when $m = 10,\ 20$ and others are same as before. In this simulation, the average of the sample sizes $\bar{k}_N$ are also computed. Table 2 shows the proportion and $\bar{k}_N$ is in the parentheses ( ). The lower bound k_{n^*} in (3.3) for known σ^2 and ρ is stated in the bottom of the Table 2.

Table 2. Accuracy of approximation and sample size

ϵ		ρ		
		0.2	0.5	0.8
0.2	$m = 10$	0.9555	0.9502	0.9497
		(38.8)	(50.3)	(64.0)
0.2	$m = 20$	0.9537	0.9506	0.9483
		(68.0)	(68.2)	(70.5)
0.4	$m = 10$	0.9566	0.9537	0.9494
		(35.0)	(44.2)	(54.7)
0.4	$m = 20$	0.9579	0.9553	0.9516
		(56.1)	(56.6)	(59.0)
	k_{n^*}	24.6	38.4	52.2

From the Table 2, the condition is satisfied in many case. When $\rho = 0.8$, the coverage probabilities are less than 0.95, but the differences from 0.95 are small. Hence the proposed two-stage procedure (3.11) is better than (3.5). The sample sizes in missing rate 0.4 are smaller than that in missing rate 0.2.

4.2. Example

We give an numerical example by using a part of the cholesterol levels for a treatment group studied at times 6, 12, 20, and 24 months, which is given in Wei and Lachin [7] or is tabulated in Srivastava [5]. We first take sample of size $m = 10$ at random from Tables 15.3.1 and 15.3.2 of Srivastava [5]. The first stage sample ($m = 10$) is in Table 3, in which * is missing.

Table 3. Cholesterol data (first stage sample)

Subject number	Months 6	12	20	24
1	268	241	260	320
2	334	290	286	320
3	313	251	307	291
4	281	277	235	210
5	252	267	299	*
6	231	285	238	251
7	279	296	262	283
8	283	248	334	271
9	272	222	246	253
10	326	304	*	*

We have $k_m = 37$, $\ell_m = 1.030$, and $u = 1426.97$ from Table 3. If we choose $d = 10$ and $\alpha = 0.05$ (hence $t = 2.262$), then $k_{N^*} = 76$ by (3.11). Next we take the second stage sample, which is in Table 4.

Table 4. Cholesterol data (second stage sample)

Subject number	Months 6	12	20	24
11	232	215	220	292
12	219	220	*	*
13	270	209	255	213
14	291	291	268	260
15	192	205	253	217
16	261	264	300	*
17	300	313	317	397
18	246	295	228	274
19	243	265	*	*
20	260	278	245	340
21	207	167	*	*
22	232	265	242	230

The sample mean is calculated as $\bar{x}_N = 265.09$. Since $t\sqrt{\ell_m u/k_N} = 9.82$, where $k_N = 78$, we have the confidence interval $[255.27,\ 274.91]$.

Acknowledgement

The authors wish to thanks to the referees for their valuable comments.

References

[1] W.C. Healy Jr., *Two-sample procedure in simultaneous estimation*, Ann. Math. Statist. **27** (1956), 687-702.

[2] H. Hyakutake, Y. Takada, and M. Aoshima, *Fixed-size confidence regions for the multinormal mean in an intraclass correlation model*, Amer. J. Math. Manage. Sci. **15** (1995), 291-308.

[3] M. Siotani, T. Hayakawa, and Y. Fujikoshi, Modern Multivariate Statistical Analysis: A Graduate Course and Handbook, American Sciences Press, 1985.

[4] M.S. Srivastava, *Estimation of intraclass correlations in familial data*, Biometrika **7** (1984), 177-185.

[5] M.S. Srivastava, Methods of Multivariate Statistics, Wiley, 2002.

[6] C. Stein, *A two-sample test for a linear hypothesis whose power is independent of the variance*, Ann. Math. Statist. **16** (1945), 245-258.

[7] L.J. Wei and J.M. Lachin, *Two-sample asymptotically distribution-free tests for incomplete multivariate observations*, J. Amer. Statist. Assoc. **79** (1984), 653-661.

Hiroto Hyakutake
Faculty of Mathematics, Kyushu University
4-2-1 Ropponmatsu, Fukuoka 810-8560, Japan
E-mail: hyakutak@math.kyushu-u.ac.jp

Masaaki Anan
Graduate School of Mathematics, Kyushu University
4-2-1 Ropponmatsu, Fukuoka 810-8560, Japan

Tomoya MIzuyoshi
Graduate School of Mathematics, Kyushu University
4-2-1 Ropponmatsu, Fukuoka 810-8560, Japan